Preventing Premature Convergence in a Simple EDA Via Global Step Size Setting

Petr Pošík

Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics
Technická 2, 166 27 Prague 6, Czech Republic
posik@labe.felk.cvut.cz

Abstract. When a simple real-valued estimation of distribution algorithm (EDA) with Gaussian model and maximum likelihood estimation of parameters is used, it converges prematurely even on the slope of the fitness function. The simplest way of preventing premature convergence by multiplying the variance estimate by a constant factor k each generation is studied. Recent works have shown that when increasing the dimensionality of the search space, such an algorithm becomes very quickly unable to traverse the slope and focus to the optimum at the same time. In this paper it is shown that when isotropic distributions with Gaussian or Cauchy distributed norms are used, the simple constant setting of k is able to ensure a reasonable behaviour of the EDA on the slope and in the valley of the fitness function at the same time.

1 Introduction

Estimation of distribution algorithms (EDAs) [1] are a class of evolutionary algorithms (EAs) that do not use the crossover and mutation operators to create the offspring population. Instead, they build a probabilistic model describing the distribution of promising individuals and create offspring by sampling from the model. In real-valued spaces, such an algorithm can have a very simple structure which is depicted in Fig. 1.

If the Gaussian distribution is employed as the model of promising individuals ([2], [3], [4], [5]), and the parameters of the distribution, μ and σ, are learned by maximum likelihood (ML) estimation, the algorithm is very prone to premature convergence (i.e. the population converges on the slope of the fitness function) as recognized by many authors (see e.g. [3], [6], [7]). In [8], it was shown also theoretically that the distance traversed by a simple Gaussian EDA with truncation selection is bounded, and [9] showed similar results for tournament selection.

Many techniques that fight the premature convergence were developed, usually by means of artificially enlarging the ML estimate of variance of the learned distribution. In [6] it is suggested to use standard deviation greater than 1 when sampling the Gaussian distribution (e.g. to use $\mathcal{G}(0, 1.5)$). Adaptive variance scaling (AVS), i.e. enlarging the variance when better solutions were found and shrinking the variance in case of no improvement, was used along with various

G. Rudolph et al. (Eds.): PPSN X, LNCS 5199, pp. 549–558, 2008.

1. Initialize the parameters $\mu^0 = (\mu_1^0, \ldots, \mu_D^0)$ and $\sigma^0 = (\sigma_1^0, \ldots, \sigma_D^0)$, D is the dimensionality of the search space. Generation counter $t = 0$.
2. Sample N offspring from the search distribution (use μ^t as the distribution center and σ^t as relative scaling factors of individual components).
3. Evaluate the individuals.
4. Select the τN best solutions (truncation selection).
5. Update parameters μ^{t+1} and σ^{t+1} using the selected individuals.
6. Enlarge the σ^{t+1} by a constant factor k (global step size).
7. Advance generation counter: $t = t + 1$.
8. If termination condition is not met, go to step 2.

Fig. 1. Simple EDA analysed in this article

techniques to trigger the AVS only on the slope of the fitness function in [10] and [11]. The algorithm in Fig. 1, that suggests enlarging the population variance by a constant factor each generation, was studied in [12] where the minimal values of the 'amplification coefficient' were determined by a search in 1D case. In [13], the theoretical model of the algorithm behavior in 1D was used to derive the minimal and maximal admissible values for k. However, in [14] it was shown experimentaly that a constant multiplier does not ensure the desired properties of the algorithm when increasing the dimensionality of the search space.

In this article it is shown that when a modified Gaussian or Cauchy distribution is used instead of the standard Gaussian distribution, the simple approach with multiplying the population variance by a constant factor ensures the desired algorithm properties. Sec. 2 introduces the requirements constituting the bounds for a reasonable behaviour of the algorithm. Sec. 3 contains description of the probability distributions compared in this article. The results of the empirical study can be found in Sec. 4 and Sec. 5 concludes the paper.

2 Fundamental Requirements on EDA

According to [15], the optimal behaviour of the self-adaptive EAs in real spaces arises from balancing two antagonistic forces: (1) the variance shrinking effect of selection, and (2) the variance enlarging effect of the variational operators (distribution sampling, in our case). In this article, an approach of [14] is used where the combined effect of the selection and variation is taken into account.

Two simple fitness landscapes are used: a linear and a sphere function:

$$f_{\text{lin}}(\mathbf{x}) = x_1 \tag{1}$$

$$f_{\text{sphere}}(\mathbf{x}) = \sum_{d=1}^{D} x_d^2 \tag{2}$$

These functions can be regarded [15] as local approximations of the real fitness functions; the fitness landscape is often modelled as consisting of slopes and

valleys (see e.g. [10], [16], [12]). The slopes and valleys are modelled with the linear (Eq. 1) and the sphere function (Eq. 2), respectively.

There are two fundamental requirements on the development of the population variance that ensure a reasonable behavior of the algorithm as a whole:

1. The variance *must not shrink on the slope*. This ensures that the population position is not bounded and that it eventually finds at least a local optimum.
2. The variance *must shrink in the valley*. In the neighborhood of the optimum, the algorithm must be allowed to converge to find the optimum precisely.

These two conditions constitute the bounds for the variance scaling factor k which must be large enough to traverse the slopes, but must not be too large to be able to focus to the optimum.

2.1 Bounds for k

The evolution of the model variance from one generation to another can be described as follows: (1) sample new individuals with variance $(\sigma^t)^2$, (2) select the best individuals, and (3) compute the variance $(\sigma^{t+1})^2$ for the next sampling. Without selection and using ML estimate, the two variances are expected to be the same. For our two fitness landscapes, the selection reduces the variance, thus

$$(\sigma^{t+1})^2 = (\sigma^t)^2 \cdot c, \tag{3}$$

where c is the ratio of the population variances in two consecutive generations, t and $t+1$, and $c < 1$ in our case. Of course, the ratio c differs for various fitness landscapes, thus it will be designated as c_{slope} and c_{valley}, respectively.

As already said in the introduction, the simplest method of preventing premature convergence is to enlarge the estimated standard deviation σ by a constant factor k (step 6 of the algorithm in Fig. 1). Thus

$$\sigma^{t+1} = k \cdot \sigma^t \cdot \sqrt{c} \tag{4}$$

In order to prevent the premature convergence on the slope, the ratio of the consecutive standard deviations should be at least 1, i.e.

$$\frac{\sigma^{t+1}}{\sigma^t} = k \cdot \sqrt{c_{\text{slope}}} \geq 1, \text{ thus } k \geq \frac{1}{\sqrt{c_{\text{slope}}}} \stackrel{\text{def}}{=} k_{\min}. \tag{5}$$

On the other hand, to be able to focus to the optimum, the model must be allowed to converge in the valley. The ratio of the two consecutive standard deviations should be lower than 1, i.e.

$$\frac{\sigma^{t+1}}{\sigma^t} = k \cdot \sqrt{c_{\text{valley}}} < 1, \text{ thus } k < \frac{1}{\sqrt{c_{\text{valley}}}} \stackrel{\text{def}}{=} k_{\max} \tag{6}$$

Joining these two conditions together gives us the bounds for the constant k:

$$k_{\min} = \frac{1}{\sqrt{c_{\text{slope}}}} \leq k < \frac{1}{\sqrt{c_{\text{valley}}}} = k_{\max} \tag{7}$$

In this paper, the value of k is called *admissible* if it satisfies condition 7.

3 Probability Distributions

Although [13] theoretically deduced bounds for k in case of 1D Gaussian distribution, in [14] it was shown that the process sketched above does not work with increasing dimensionality, since the interval of admissible k diminishes and eventually vanishes. This is due to the fact that the variance after selection in the neighborhood of the valley (sphere function) increases with dimensionality, thus $k_{\max}$ must be successively smaller and eventually gets lower than $k_{\min}$.

In this article, a distribution that does not exhibit this unpleasant behaviour is sought for. Three distributions are compared.

Standard Gaussian distribution (designated as $\mathcal{G}$)**.** Probably the most often used distribution in real-valued evolutionary algorithms. The 1D normal distribution with zero mean and variance σ^2 has the following p.d.f.:

$$f_{\mathcal{N}(0,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} \tag{8}$$

Sampling process. D-dimensional realizations of the standard normal distribution can be created by sampling each component independently from the 1D standard normal distribution.

Isotropic distribution with 1D Gaussian norm (designated as $\mathcal{G}^{\text{iso}}$)**.**[1] Used in the hope that it preserves some features of the 1D Gaussian distribution.

Sampling process. 1D version of $\mathcal{G}^{\text{iso}}$ is the same as 1D version $\mathcal{G}$. The multidimensional versions of $\mathcal{G}^{\text{iso}}$ can be created by (1) sampling the direction vector uniformly on the unit hypersphere[2], and (2) by multiplying the vector by a factor sampled from χ-distribution with 1 degree of freedeom. The χ-distribution describes norms of vectors generated from $\mathcal{G}$.

Isotropic distribution with 1D Cauchy norm (designated as $\mathcal{C}^{\text{iso}}$)**.** Selected for the comparison to show the effects of heavy tails (if any). The 1D Cauchy distribution with median 0 and upper quartile γ has the following p.d.f.:

$$f_{\mathcal{C}(0,\gamma)} = \frac{1}{\pi} \frac{\gamma}{x^2 + \gamma^2} \tag{9}$$

Standard Cauchy distributed values with $\gamma = 1$ can be be obtained by sampling two values from $\mathcal{G}$ and dividing them.

Sampling process. D-dimensional realizations of $\mathcal{C}^{\text{iso}}$ can be sampled similarly as $\mathcal{G}^{\text{iso}}$ with the exception that the multiplication factor is sampled from 1D Cauchy distribution instead of 1D Gaussian.

[1] In this article, the term *isotropic* is not meant as a feature of the distribution (standard normal distribution is isotropic as well); it describes the sampling process.

[2] Sampling a vector on a unit hypersphere can be achieved e.g. by sampling D-dimensional standard normal distribution and dividing the resulting vector by its norm.

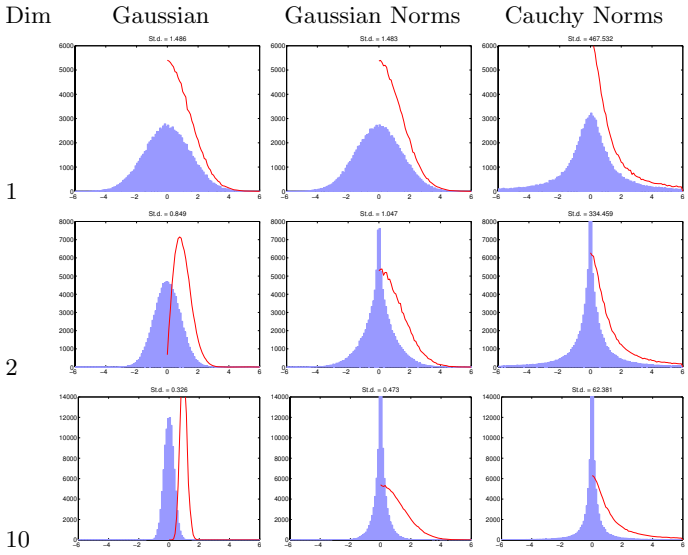

Fig. 2. The distribution of the first coordinate (the histograms) and the distribution of the vector norms (solid line) for $\mathcal{G}$, $\mathcal{G}^{\text{iso}}$, and $\mathcal{C}^{\text{iso}}$, and for the search space dimensions 1, 2, and 10

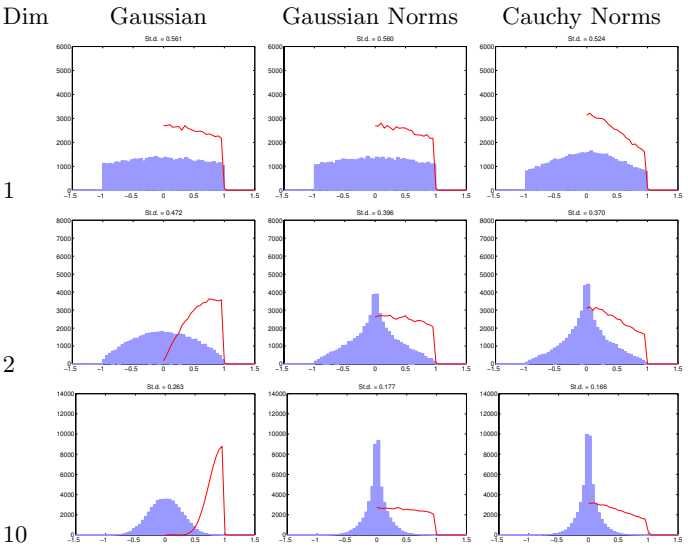

Fig. 3. After selection with sphere function. The distribution of the first coordinate (the histograms) and the distribution of the vector norms (solid line) for $\mathcal{G}$, $\mathcal{G}^{\text{iso}}$, and $\mathcal{C}^{\text{iso}}$, and for the search space dimensions 1, 2, and 10. Note that the distribution of vector norms (solid line) is cut off at $x = 1$ due to the modification of sampling process described in Sec. 3.1.

These distributions were already studied in several works from different points of view. In [17], the local convergence rates of evolutionary algorithms with Gaussian and Cauchy mutations are estimated and compared. In [18], the convergence to a local optimum was studied as well, along with the ability to locate narrow valleys and the influence of the dimensionality on the exploration efficiency. The usefulness of the Cauchy distributions in case of multimodal optimization was explored in [19]. In this article it is studied if these distributions allow for the simple constant setting of the global step size.

3.1 Modification of Vector Norms

It was deliberately decided to normalize[3] the vector norms of all three distributions in such a way, that the 100τ-percentile of the distribution of norms equals to 1. This is achieved simply by

- dividing the $\mathcal{G}$-distributed vectors by the value of inverse cumulative distribution function (i.c.d.f) of the χ distribution with D degrees of freedom (d.o.f.) at point τ, i.e. $\mathbf{x}_m = \mathbf{x}/CDF^{-1}_{\chi_D}(\tau)$, $\mathbf{x} \sim \mathcal{G}$,
- dividing the $\mathcal{G}^{\text{iso}}$-distributed vectors by the value of the i.c.d.f. of the χ distribution with 1 d.o.f. at point τ, i.e. $\mathbf{x}_m = \mathbf{x}/CDF^{-1}_{\chi_1}(\tau)$, $\mathbf{x} \sim \mathcal{G}^{\text{iso}}$, or by
- dividing the $\mathcal{C}^{\text{iso}}$-distributed vectors by the value of the i.c.d.f. of the standard Cauchy distribution at point $(1+\tau)/2$, i.e. $\mathbf{x}_m = \mathbf{x}/CDF^{-1}_{\mathcal{C}}(\frac{1+\tau}{2})$, $\mathbf{x} \sim \mathcal{C}^{\text{iso}}$, respectively.

The distributions of sampled data points and their norms are depicted in Fig. 2. The fact that the 100τ-percentile of the norm distribution is equal to 1 is demonstrated in Fig. 3 which shows the distributions of selected data points when sphere function is used. The frequency of norms of the selected data points is cut off at value 1.

4 Experiments, Results and Discussion

The bounds for k for all three distributions were found experimentaly. The lower bound $k_{\min}$ is found by using the f_{lin}, the upper bound $k_{\max}$ is found by experiments with f_{sphere}. During the experiments, the value of standard deviation of coordinate x_1 is tracked and it is checked if it increases or decreases (on average). The bisection method is used to determine the value of k for which the variance stays the same (with certain tolerance).

The population size 1,000 was used in all experiments. To determine each particular $k_{\min}$ (and $k_{\max}$), 10 independent runs of 100 generations were carried out. Each run was started with initial parameters $\mu^0 = 0$ and $\sigma^0 = 1$ ensuring that the processes are started in the stationary state. During each run, the standard deviation of x_1 was tracked; this gives 10 values of st.d. for each of 100

[3] There is no special need for the normalization. With the normalization, however, the graphs in Figs. 4 and 5 show more regular patterns and are more comparable.

generations. To this data, a linear function of the form $E(\log(\text{st.d.})) = a \cdot \text{gen} + b$ was fitted ('gen' is the generation counter) using simple linear regression which should be adequate type of model. The sign of the learned parameter a was used to decide, if the variances increase or decrease during the run.

The bounds of k found for $\mathcal{G}$ can be seen in Fig. 4. For 1D search space there exists an interval of admissible values of k for all tested selection proportions τ. However, with increasing dimensionality, the value of $k_{\min}$ grows faster than $k_{\max}$ for all values of τ, and for dimensions greater then 5 there is no admissible k (which would ensure effective traversing of slopes and focusing to the optimum in the same time). This is in accordance with the results in [13] and [14].

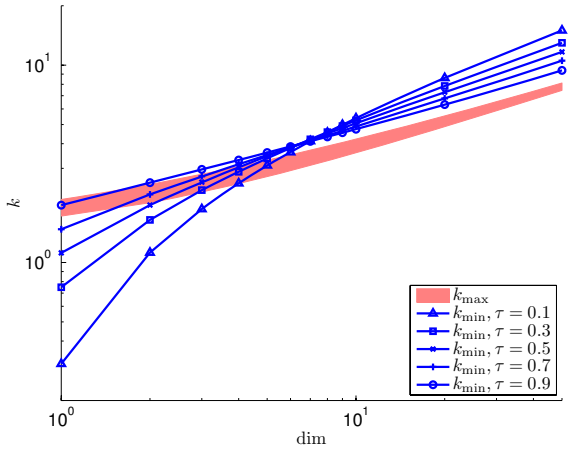

Fig. 4. Minimal and maximal values of k for the Gaussian distribution. (The lines for the respective $k_{\max}$ are very close to each other; without losing the big picture they were replaced by the shaded region.) It can be observed that for $D > 5$ the $k_{\min}$ is greater than $k_{\max}$ for all tested selection proportions τ and the admissible interval for k does not exist!

The same figures when $\mathcal{G}^{\text{iso}}$ is used are depicted in Fig. 5, left. The results are completely different now! For all but the highest values of τ, there seems to exist an interval of admisible values of k and this interval does not shrink with incresing dimensionality.

The situation for $\mathcal{C}^{\text{iso}}$ distribution is even better, see Fig. 5, right. The size of admissible interval for k does not shrink so much when increasing the selection proportion τ, as was the case for $\mathcal{G}^{\text{iso}}$.

It can be also observed that for the isotropic distributions and a particular value of selection proportion τ, the ratio $k_{\max}/k_{\min}$ stays almost the same regardless of the dimensionality. This observation could be used to create a simple equation for the setting of k in relation to τ and the dimensionality. Of course, optimal setting of k depends on the problem, on the initial values of μ^0 and σ^0, and can also depend on the search distribution used. At this moment, it is not

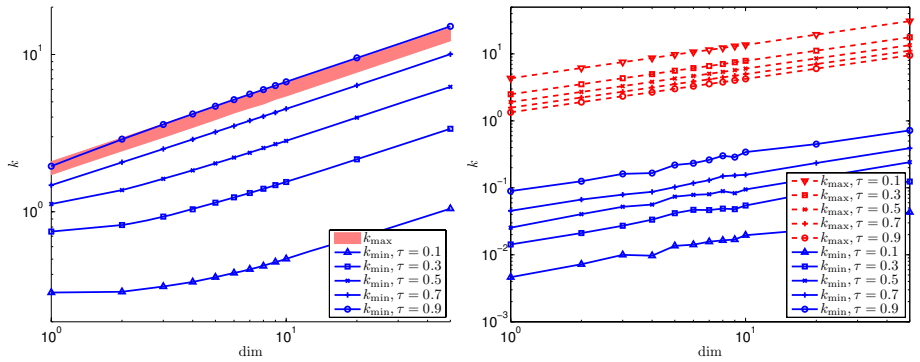

Fig. 5. Minimal and maximal values of k for the isotropic Gaussian (on the left) and isotropic Cauchy (on the right) distributions. (The lines for the respective $k_{\max}$ of the $\mathcal{G}^{\text{iso}}$ distribution are very close to each other; without losing the big picture they were replaced by the shaded region.) It can be observed that the admissible interval for k exists and does not shrink with the dimensionality for almost all tested selection proportions τ and for both tested isotropic distributions.

clear if it is better "on average" to set k only slightly above $k_{\min}$, slightly below $k_{\max}$, or somewhere in the middle.

As already said in the introduction, in [6] the authors showed that their EDA with truncation selection with $\tau = 0.3$ which used the value of 1.5 for the standard deviation of the Gaussian distribution was able to find the optimum of the 10D Rosenbrock function while EDA without this modification (using ML estimate of σ) converged prematurely. The value 1.5 can be transformed to the context of this article; the corresponding $k = 1.5 \cdot CDF^{-1}_{\chi_{10}}(0.3) \approx 4$. Looking at the Fig. 4 (dim=10, $\tau = 0.3$) we can see that this value is not admissible; it lies somewhere in the shaded region of $k_{\max}$, below $k_{\min}$. Thus, the population variance was shrinking during the whole evolution (as shown in [6]). The shrinking was a bit slower, however, than when using ML estimate of σ giving the algorithm the time needed to find the global optimum. The algorithm was started from the origin. If it was started from a more distant point, the results obtained in this article suggest that the optimum would not be reached.

The adaptive variance scaling approach (AVS) presented in [10] and [11] should work even for the isotropic distributions used in this article. Since it is a dynamic scheme for setting the k, it needn't be limited to admissible values of k. For the algorithm it is often profitable to set $k > k_{\max}$ when on slope, or to set $k < k_{\min}$ when in the valley which ensures faster traversal of slopes and faster convergence to the optimum, respectively. On the other hand, AVS alone is an iterative update scheme and it can take several generations to switch the scaling from slope-style to valley-style or vice versa. That is the reason behind the triggers introduced in [10] and [11] which should decide if the population is on the slope or in the valley and trigger the AVS only on the slope; in the valley, the ML estimate of σ is used without scaling. The right behavior of such an algorithm is largely determined by the ability of the trigger to decide correctly

whether to trigger the scaling. The results of this article can thus be useful for these algorithms in two ways: (1) if the trigger is good, the scaling factor can be set to at least $k_{\min}$ on the slope, and at most to $k_{\max}$ in the valley, or (2) if the trigger makes mistakes, the algorithm can use the admissible interval of $\langle k_{\min}, k_{\max}\rangle$ as a safeguard.

5 Summary and Future Work

This article aimed at simple way of preventing premature convergence of a simple EDA. The variance of the distribution estimated from the selected data is increased by the factor (or global step size) k each generation, artificially keeping the sufficient diversity in the population.

Recent works have shown that when Gaussian distribution is used, a constant value of k which would ensure a reasonable behaviour of the algorithm on the slopes *and* in the valleys of the fitness function exists only for low-dimensional spaces.

The situation is much better when isotropic distribution with Gaussian or Cauchy norms is used. Both of these two distributions ensure the existence of the admissible interval for k for a broad range of selection proportions τ and search space dimensionalities. Moreover, the ratio $k_{\max}/k_{\min}$ stays almost the same for the isotropic distributions, with Cauchy distribution giving larger margin.

Compiling a practically appliable heuristic for setting the value of k, building a real working optimization algorithm based on these principles, and its comparison with other scaling techniques remain as the future work. It would be also appealing to explore this technique in combination with other selection schemes different from the truncation selection.

Acknowledgements

The project was supported by the Ministry of Education, Youth and Sport of the Czech Republic with the grant No. MSM6840770012 entitled "Transdisciplinary Research in Biomedical Engineering II".

References

1. Larrañaga, P., Lozano, J.A. (eds.): Estimation of Distribution Algorithms. GENA. Kluwer Academic Publishers, Dordrecht (2002)
2. Larrañaga, P., Lozano, J.A., Bengoetxea, E.: Estimation of distribution algorithms based on multivariate normal distributions and gaussian networks. Technical Report KZZA-IK-1-01, Dept. of Computer Science and Artificial Intelligence, University of Basque Country (2001)
3. Bosman, P.A.N., Thierens, D.: Expanding from discrete to continuous estimation of distribution algorithms: The IDEA. In: PPSN VI: Proceedings of the 6th International Conference on Parallel Problem Solving from Nature, London, UK, pp. 767–776. Springer, Heidelberg (2000)
4. Rudlof, S., Köppen, M.: Stochastic hill climbing by vectors of normal distributions. In: First Online Workshop on Soft Computing, Nagoya, Japan (1996)

5. Ahn, C.W., Ramakrishna, R.S., Goldberg, D.E.: Real-coded bayesian optimization algorithm: Bringing the strength of BOA into the continuous world. In: Deb, K. (ed.) Proceedings of the Genetic and Evolutionary Computation Conference GECCO 2004, pp. 840–851. Springer, Heidelberg (2004)
6. Yuan, B., Gallagher, M.: On the importance of diversity maintenance in estimation of distribution algorithms. In: Beyer, H.G., O'Reilly, U.M. (eds.) Proceedings of the Genetic and Evolutionary Computation Conference GECCO 2005, vol. 1, pp. 719–726. ACM Press, New York (2005)
7. Ocenasek, J., Kern, S., Hansen, N., Koumoutsakos, P.: A mixed bayesian optimization algorithm with variance adaptation. In: Yao, X., Burke, E.K., Lozano, J.A., Smith, J., Merelo-Guervós, J.J., Bullinaria, J.A., Rowe, J.E., Tiňo, P., Kabán, A., Schwefel, H.-P. (eds.) PPSN 2004. LNCS, vol. 3242, pp. 352–361. Springer, Heidelberg (2004)
8. Grahl, J., Minner, S., Rothlauf, F.: Behaviour of UMDAc with truncation selection on monotonous functions. In: IEEE Congress on Evolutionary Computation, CEC 2005, vol. 3, pp. 2553–2559 (2005)
9. Gonzales, C., Lozano, J., Larranaga, P.: Mathematical modelling of UMDAc algorithm with tournament selection. International Journal of Approximate Reasoning 31(3), 313–340 (2002)
10. Grahl, J., Bosman, P.A.N., Rothlauf, F.: The correlation-triggered adaptive variance scaling IDEA. In: Proceedings of the 8th annual conference on Genetic and Evolutionary Computation Conference - GECCO 2006, pp. 397–404. ACM Press, New York (2006)
11. Bosman, P.A.N., Grahl, J., Rothlauf, F.: SDR: A better trigger for adaptive variance scaling in normal EDAs. In: GECCO 2007: Proceedings of the 9th annual conference on Genetic and Evolutionary Computation, pp. 492–499. ACM Press, New York (2007)
12. Yuan, B., Gallagher, M.: A mathematical modelling technique for the analysis of the dynamics of a simple continuous EDA. In: IEEE Congress on Evolutionary Computation, CEC 2006, Vancouver, Canada, pp. 1585–1591. IEEE Press, Los Alamitos (2006)
13. Pošík, P.: Gaussian EDA and truncation selection: Setting limits for sustainable progress. In: IEEE SMC International Conference on Distributed Human-Machine Systems, DHMS 2008, Athens, Greece. IEEE, Los Alamitos (2008)
14. Pošík, P.: Truncation selection and gaussian EDA: Bounds for sustainable progress in high-dimensional spaces. In: Giacobini, M., Brabazon, A., Cagnoni, S., Di Caro, G.A., Drechsler, R., Ekárt, A., Esparcia-Alcázar, A.I., Farooq, M., Fink, A., McCormack, J., O'Neill, M., Romero, J., Rothlauf, F., Squillero, G., Uyar, A.Ş., Yang, S. (eds.) EvoWorkshops 2008. LNCS, vol. 4974, pp. 525–534. Springer, Heidelberg (2008)
15. Beyer, H.G., Deb, K.: On self-adaptive features in real-parameter evolutionary algorithms. IEEE Trans. on Evol. Comp. 5(3), 250–270 (2001)
16. Grahl, J., Bosman, P.A.N., Minner, S.: Convergence phases, variance trajectories, and runtime analysis of continuous EDAs. In: GECCO 2007: Proceedings of the 9th annual conference on Genetic and evolutionary computation, pp. 516–522. ACM Press, New York (2007)
17. Rudolph, G.: Local convergence rates of simple evolutionary algorithms with cauchy mutations. IEEE Transactions on Evolutionary Computation 1, 249–258 (1997)
18. Obuchowicz, A.: Multidimensional mutations in evolutionary algorithms based on real-valued representation. Int. J. Systems Science 34(7), 469–483 (2003)
19. Hansen, N., Gemperle, F., Auger, A., Koumoutsakos, P.: When do heavy-tail distributions help? In: Parallel Problem Solving from Nature - PPSN IX, pp. 62–71. Springer, Heidelberg (2006)